Visual geo-localization of non-photographic depictions via 2D-3D alignment

Mathieu Aubry, Bryan C. Russell, Josef Sivic

Visual geo-localization of non-photographic depictions via 2D-3D alignment

Mathieu Aubry, Bryan Russell and Josef Sivic

Abstract This chapter describes a technique that can geo-localize arbitrary 2D depictions of architectural sites, including drawings, paintings and historical photographs. This is achieved by aligning the input depiction with a 3D model of the corresponding site. The task is very difficult as the appearance and scene structure in the 2D depictions can be very different from the appearance and geometry of the 3D model, e.g., due to the specific rendering style, drawing error, age, lighting or change of seasons. In addition, we face a hard search problem: the number of possible alignments of the depiction to a set of 3D models from different architectural sites is huge. To address these issues, we develop a compact representation of complex 3D scenes. 3D models of several scenes are represented by a set of discriminative visual elements that are automatically learnt from rendered views. Similar to object detection, the set of visual elements, as well as the weights of individual features for each element, are learnt in a discriminative fashion. We show that the learnt visual elements are reliably matched in 2D depictions of the scene despite large variations in rendering style (e.g. watercolor, sketch, historical photograph) and structural changes (e.g. missing scene parts, large occluders) of the scene. We demonstrate that the proposed approach can automatically identify the correct architectural site as well as recover an approximate viewpoint of historical photographs and paintings with respect to the 3D model of the site.

Mathieu Aubry
Inria / Université Paris-Est, LIGM (UMR CNRS 8049), ENPC, F-77455 Marne-la-Vallée e-mail: Mathieu.Aubry@imagine.enpc.fr

Bryan Russell
Adobe Research e-mail: brussell@adobe.com

Josef Sivic
Inria, WILLOW project-team, Département d'Informatique de l'Ecole Normale Supérieure, ENS/INRIA/CNRS UMR 8548 e-mail: Josef.Sivic@ens.fr

Fig. 1 Our system automatically geo-localizes paintings, drawings, and historical photographs by recovering their viewpoint with respect to a geo-referenced 3D model of the depicted architectural site. Here geo-localized paintings of Notre Dame in Paris are visualized in the Google Earth geobrowser.

1 Introduction

In this work we seek to automatically geo-localize historical photographs and non-photographic renderings, such as paintings and line drawings, by matching them with a set of geo-referenced 3D models. Specifically, we wish to establish a set of point correspondences between local structures on the 3D models and their respective 2D depictions. The established correspondences will in turn allow us to identify the correct architectural site and find an approximate viewpoint of the 2D depiction with respect to the identified 3D model, thus geo-localizing the input depiction. We focus on depictions that are, at least approximately, perspective renderings of the 3D scene. Example results are shown in figure 1. We show that our alignment method works with complex textured 3D models obtained by recent multi-view stereo reconstruction systems [22] as well as with simplified models obtained from 3D modeling tools such as Trimble 3D Warehouse that often appear in geobrowsing tools such as Google Earth.

Why is this task important? First, non-photographic depictions are plentiful and comprise a large portion of our visual record. We wish to reason about them, and aligning such depictions to our 3D physical world is an important step towards this goal. Second, such depictions are often stored in archives and museums with limited access and search capabilities. Automatic large-scale geo-localization would change the way archivists access and organize such imagery. Finally, reliable auto-

matic image-to-3D model matching is important in domains where geo-referenced 3D models are often available, but may contain errors or unexpected changes (e.g. something built/destroyed) [7], such as urban planning, civil engineering or archaeology.

The task of aligning 3D models to 2D non-photographic depictions is extremely challenging. As discussed in prior work [41, 47], local feature matching based on interest points (e.g. SIFT [35]) often fails to find correspondences across paintings and photographs. First, the rendering styles across the two domains can vary considerably. The scene appearance (colors, lighting, texture) and geometry depicted by the artist can be very different from the rendering of the 3D model, e.g. due to the depiction style, drawing error, or changes in the geometry of the scene. Second, we face a hard search problem. The number of possible alignments of the depiction to a large 3D model, such as a partial reconstruction of a city, is huge. Which parts of the depiction should be aligned to which parts of the 3D model? How does one search over the possible alignments?

To address these issues we introduce the idea of automatically discovering *discriminative visual elements* for a 3D scene. We define a discriminative visual element to be a mid-level patch that is rendered with respect to a given viewpoint from a 3D model with the following properties: (i) it is visually discriminative with respect to the rest of the "visual world" represented here by a generic set of randomly sampled patches, (ii) it is distinctive with respect to other patches in nearby views, and (iii) it can be reliably matched across nearby viewpoints. We employ modern representations and recent methods for discriminative learning of visual appearance, which have been successfully used in recent object recognition systems. Our method can be viewed as "multi-view geometry [27] meets part-based object recognition [18]" – here we wish to automatically discover the distinctive object parts for a large 3D site.

We discover discriminative visual elements by first sampling candidate mid-level patches across different rendered views of the 3D model. We cast the image matching problem as a classification task over appearance features with the candidate mid-level patch as a single positive example and a negative set consisting of a large set of "background" patches. Note that a similar idea has been used in learning per-exemplar distances [20] or per-exemplar support vector machine (SVM) classifiers [36] for object recognition and cross-domain image retrieval [47]. Here we apply per-exemplar learning for matching mid-level structures between images.

For a candidate mid-level patch to be considered a discriminative visual element, we require that (i) it has a low training error when learning the matching classifier, and (ii) it is reliably detectable in nearby views via cross-validation. Critical to the success of operationalizing the above procedure is the ability to efficiently train linear classifiers over Histogram of Oriented Gradients (HOG) features [13] for each candidate mid-level patch, which has potentially millions of negative training examples. In contrast to training a separate SVM classifier for each mid-level patch, we change the loss to a square loss, similar to [5, 23]. We show that the solution can be computed in closed form, which is computationally more efficient as it does not require expensive iterative training. In turn, we show that efficient train-

ing opens-up the possibility to evaluate the discriminability of millions of candidate visual elements densely sampled over all the rendered views. We further show how our formulation is related to recent work that performs linear discriminant analysis (LDA) by analyzing a large set of negative training examples and recovering the sample mean and covariance matrix that decorrelates the HOG features [26, 23].

The output for each discriminative visual element is a trained classifier. At runtime, for an input depiction (e.g. a painting), we run the set of trained classifiers in a sliding-window fashion across different scales. Detections with high responses are considered as putative correspondences with the 3D model, from which camera resectioning is performed. The output is a geo-localization of the input depiction in the form of its approximate viewpoint with respect to the georeferenced 3D model. We show that our approach is able to scale to a number of different 3D sites and handles different input rendering styles. To evaluate our alignment procedure, we use the publicly available dataset of [2]. First, we evaluate whether the proposed technique can coarsely localize the input depiction by correctly identifying the 3D model corresponding to the depicted architectural site. Second, for the correctly coarsely localized depictions we perform a user study where human subjects are asked to judge the goodness of the output alignment. Parts of this chapter were previously published in [2]. Here we apply the 2D-to-3D alignment technique described in [2] to the task of automatic geo-localization of historical and non-photographic imagery.

2 Related work

This section reviews prior work on visual geo-localization with a focus on non-photographic and historical imagery.

Visual geo-localization using local features. Local invariant features and descriptors such as SIFT [35] represent a powerful tool for matching photographs of the same at least lightly textured scene despite changes in viewpoint, scale, illumination, and partial occlusion. Without explicitly representing the 3D structure of the scene, visual geo-localization can be cast as large-scale instance-level retrieval [37, 39, 48]. Local invariant features are extracted from each image in a geo-referenced image database. The query photograph is then localized despite changes in viewpoint or illumination by finding the best matching image in the database and transferring its geotag [8, 12, 25, 31, 43, 50, 51]. Large 3D scenes, such as a portion of a city [33], can be also represented as a geo-referenced 3D point cloud with associated local feature descriptors extracted from the corresponding photographs [30, 33, 42]. Geo-referenced camera pose of a given query photograph can be recovered from 2D to 3D correspondences obtained by matching appearance of local features verified using geometric constraints [27]. However, appearance changes beyond the modeled invariance, such as significant perspective distortions, non-rigid deformations, non-linear illumination changes (e.g. shadows), weathering, change of seasons, structural variations or a different depiction style (photograph,

painting, sketch, drawing) cause local feature-based methods to fail [28, 41, 47]. Greater insensitivity to appearance variation can be achieved by matching the geometric or symmetry pattern of local image features [9, 28, 46], rather than the local features themselves. However, such patterns have to be detectable and consistent between the matched views.

Visual geo-localization via alignment of contours. Contour-based 2D to 3D alignment methods [29, 34] rely on detecting edges in the image and aligning them with projected 3D model contours. Such approaches are successful if scene contours can be reliably extracted both from the 2D image and the 3D model. A recent example is the work on photograph localization using semi-automatically extracted skylines matched to clean contours obtained from rendered views of digital elevation models [3, 4]. Contour matching was also used for aligning paintings to 3D meshes reconstructed from photographs [41]. However, contours extracted from paintings and real-world 3D meshes obtained from photographs are noisy. As a result, the method requires a good initialization with a close-by viewpoint. In general, reliable contour extraction is a hard and yet unsolved problem.

Visual geo-localization with discriminative image representations. Modern image representations developed for visual recognition, such as HOG descriptors [13], represent 2D views of objects or object parts [18] by a weighted spatial distribution of image gradient orientations. The weights are learnt in a discriminative fashion to emphasize object contours and de-emphasize non-object, background contours and clutter. Such a representation can capture complex object boundaries in a soft manner, avoiding hard decisions about the presence and connectivity of imaged object edges. Learnt weights have also been shown to emphasize visually salient image structures matchable across different image domains, and have been used to coarsely geo-localize non-photographic depictions such as paintings or sketches using a global image descriptor [47]. Similar representation has been used to learn architectural elements that summarize a certain geo-spatial area by analyzing (approximately rectified) 2D street-view photographs from multiple cities [15] and to detect objects depicted in paintings which have been trained from images [11].

Building on discriminatively-trained models for object detection, we develop a compact representation of 3D scenes suitable for alignment and visual geo-localization of arbitrary 2D depictions, such as paintings, drawings, or historical photographs. In contrast to [15, 47] who analyze 2D images, our method takes advantage of the knowledge and control over the 3D model to learn a set of mid-level 3D scene elements robust to a certain amount of viewpoint variation and capable of recovery of the (approximate) geo-referenced camera viewpoint. We show that the learnt mid-level scene elements are reliably detectable in 2D depictions of the scene despite large changes in appearance and rendering style.

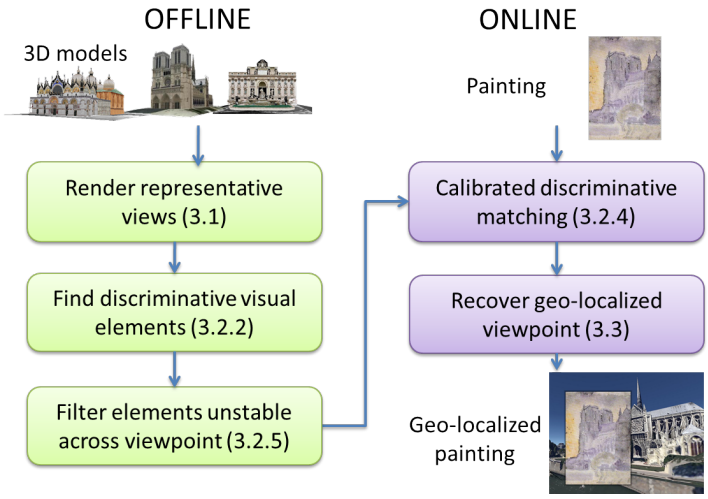


Fig. 2 Approach overview. In the offline stage (left) we summarize a set of given geo-referenced 3D models using a collection of discriminative visual elements. In the online stage (right) we match the learnt visual elements to the input depiction and use the obtained correspondences to recover the camera viewpoint with respect to the best matching 3D model.

3 Geo-localization by matching discriminative visual elements

The proposed method has two stages: first, in an offline stage we learn a set of discriminative visual elements representing one or more architectural sites; second, in an online stage a given unseen query depiction is aligned with the appropriate 3D model by matching with the learnt visual elements. The proposed algorithm is summarized in figure 2. In detail, the input to the offline stage are 3D models of multiple architectural sites. The output is a set of view-dependent visual element detectors able to identify specific structures of the different 3D models in various types of 2D imagery. The approach begins by rendering a set of representative views of each 3D model. Next, a set of visual element detectors is computed from the rendered views by identifying scene parts that are discriminative and can be reliably detected over a range of viewpoints. During the online stage, given an input 2D depiction, we match with the learnt visual element detectors and use the top scoring detections to recover a camera viewpoint with respect to the best matching 3D model.

3.1 Rendering representative views

We sample possible views of each 3D model in a similar manner to [3, 30, 41]. First, we identify the ground plane and corresponding vertical direction. The camera positions are then sampled on the ground plane on a regular grid. For each camera

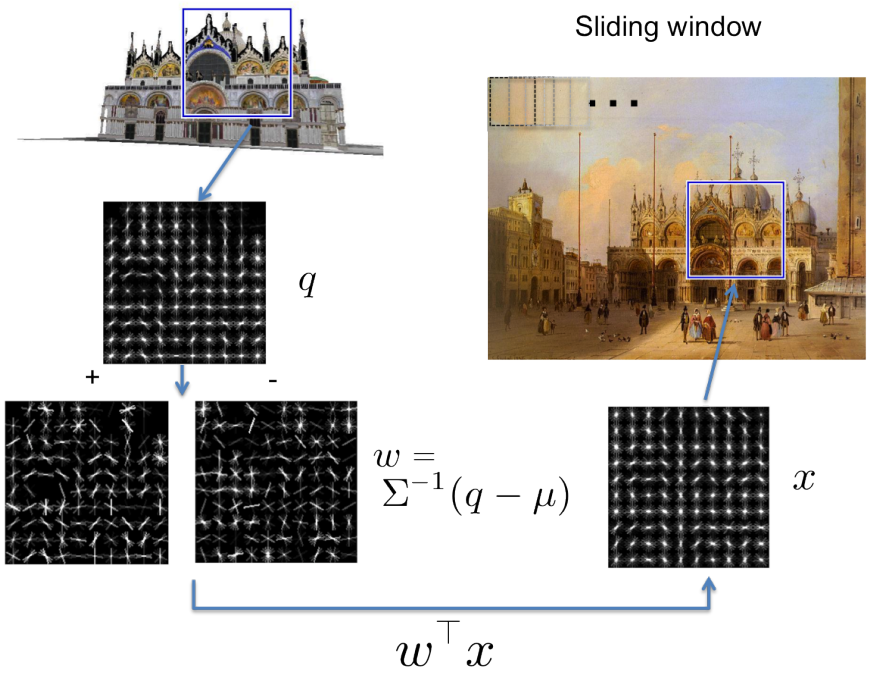


Fig. 3 Matching as classification. Given a region and its HOG descriptor q in a rendered view (top left) the aim is to find the corresponding region in a depiction (e.g. a painting, top right). This is achieved by training a linear HOG-based sliding window classifier using q as a single positive example and a large number of negative data. The classifier weight vector w is visualized by separately showing the positive (+) and negative (-) weights at different orientations and spatial locations. The best match x in the depiction is found as the maximum of the classification score.

position we sample 12 possible horizontal camera rotations assuming no in-plane rotation of the camera. For each horizontal rotation we sample 2 vertical rotations (pitch angles). Views where less than 5% of the pixels are occupied by the 3D model are discarded. This procedure results in 7,000-45,000 views for each model depending on the size of the 3D site. Note that the rendered views form only an intermediate representation and can be discarded after visual element detectors are extracted.

3.2 Finding and matching discriminative visual elements

3.2.1 Matching as classification

The aim is to match a given rectangular image patch q (represented by a HOG descriptor [13]) in a rendered view to its corresponding image patch in the depiction, as

illustrated in figure 3. Instead of finding the best match measured by the Euclidean distance between the descriptors, we train a linear classifier with q as a single positive example (with label $y_q = +1$) and a large number of negative examples x_i for $i = 1$ to N (with labels $y_i = -1$). The matching is then performed by finding the patch x^* in the depiction with the highest classification score

$$s(x) = w^\top x + b, \tag{1}$$

where w and b are the parameters of the linear classifier.

Parameters w and b can be obtained by minimizing a cost function of the following form

$$E(w,b) = L\left(y_q, w^T q + b\right) + \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, w^T x_i + b\right), \tag{2}$$

where the first term measures the loss L on the positive example q (also called "exemplar") and the second term measures the loss on the negative data. A regularizer could be added to this cost E, but we found that was not necessary with our choice of loss functions. A particular case of the exemplar-based classifier is the exemplar-SVM [36, 47], where the loss $L(y, s(x))$ between the label y and predicted score $s(x)$ is the hinge-loss $L(y, s(x)) = \max\{0, 1 - ys(x)\}$ [6]. For exemplar-SVM cost (2) is convex and can be minimized using iterative algorithms [17, 45], but this remains computationally expensive.

3.2.2 Selection of discriminative visual elements via least squares regression

Using instead a square loss $L(y, s(x)) = (y - s(x))^2$, similarly to [5, 23], w_{LS} and b_{LS} minimizing (2) and the optimal cost E^*_{LS} can be obtained in closed form as

$$w_{LS} = \frac{2}{2 + \|\Phi(q)\|^2}\Sigma^{-1}(q - \mu), \tag{3}$$

$$b_{LS} = -\frac{1}{2}(q + \mu)^T w_{LS}, \tag{4}$$

$$E^*_{LS} = \frac{4}{2 + \|\Phi(q)\|^2}, \tag{5}$$

where $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ denotes the mean of the negative examples, $\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^\top$ their covariance and Φ is the "whitening" transformation such that

$$\|\Phi(x)\|^2 = (x - \mu)^\top \Sigma^{-1}(x - \mu). \tag{6}$$

We can use the value of the optimal cost (5) as a measure of the discriminability of a specific q. If the training cost (error) for a specific candidate visual element q is small, this visual element can be easily separated from the negative data and thus it is discriminative. This observation can be translated into a simple and efficient

algorithm for ranking candidate element detectors based on their discriminability. Given a rendered view, we consider as candidate visual elements all patches that are local minima (in scale and space) of the training cost (5)

3.2.3 Relation to linear discriminant analysis (LDA)

An alternative way to compute w and b is to use LDA, similarly to [23, 26]. This results in slightly different values of the parameters:

$$w_{LDA} = \Sigma^{-1}(q - \mu_n), \tag{7}$$

and

$$b_{LDA} = \frac{1}{2}\left(\mu^T \Sigma^{-1} \mu - q^T \Sigma^{-1} q\right). \tag{8}$$

Classifiers obtained by minimizing the least squares cost function (2) or satisfying the LDA ratio test can be used for matching a candidate visual element q to a 2D depiction as described in equation (1). Note that the decision hyperplanes obtained from the least squares regression, w_{LS}, and linear discriminant analysis, w_{LDA}, are parallel. As a consequence, for a particular visual element q the ranking of matches according to the matching score (1) would be identical for the two methods. In other words, in an object detection setup [13, 26, 23] the two methods would produce identical precision-recall curves. In our matching setup, for a given q the best match in a particular depiction would be identical for both methods. The actual value of the score, however, becomes important when comparing matching scores across different visual element detectors q. In object detection, the score of the learnt classifiers is typically calibrated on a held-out set of labeled validation examples [36].

3.2.4 Calibrated discriminative matching

We have found that calibration of matching scores across different visual elements is important for the quality of the final matching results. Below we describe a procedure to calibrate matching scores without the need of any labelled data. First, we found [2] that the matching score obtained from LDA produces significantly better matching results than matching via least squares regression. Nevertheless, we found that the raw uncalibrated LDA score favors low-contrast image regions, which have an almost zero HOG descriptor. To avoid this problem, we further calibrate the LDA score by subtracting a term that measures the score of the visual element q matched to a low-contrast region, represented by zero (empty) HOG vector

$$s_{calib}(x) = s_{LDA}(x) - s_{LDA}(0) \tag{9}$$

$$= (q - \mu)^T \Sigma^{-1} x. \tag{10}$$

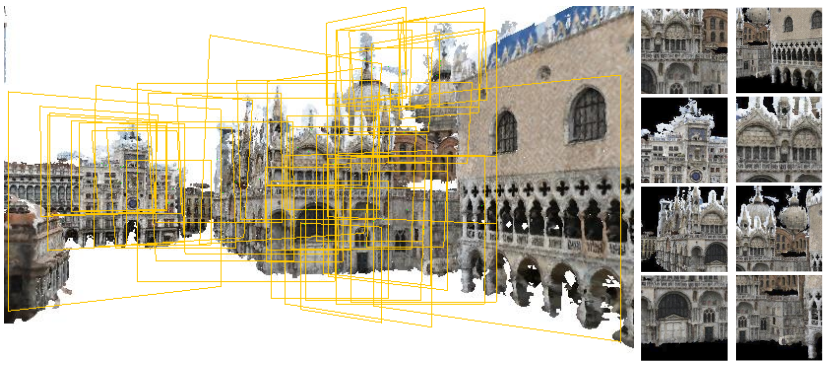


Fig. 4 Examples of selected visual elements for a 3D site. Left: Selection of top ranked 50 visual elements visible from this specific view of the site. Each element is depicted as a planar patch with an orientation of the plane parallel to the camera plane of its corresponding source view. **Right:** Subset of 8 elements shown from their original viewpoints. Note that the proposed algorithm prefers visually salient scene structures such as the two towers in the top-right or the building in the left part of the view.

This calibrated score gives much better results on the dataset of [28], as shown in [2], and significantly improves matching results.

3.2.5 Filtering elements unstable across viewpoint

We discard elements that cannot be reliably detected in close-by rendered views. This filtering criterion removes many unstable elements that are, for example, ambiguous because of repeated structures in the rendered view or cover large depth discontinuities and hence significantly change with viewpoint. To achieve that, we perform two additional tests on each visual element. First, to suppress potential repeated structures, we require that the ratio between the score of the first and second highest scoring detection in the image is larger than a threshold of 1.04, similar to [35]. Second, we run the discriminative elements in the views near the one where they were defined and keep only visual elements that are successfully detected in more than 80% of the nearby views. The definition of what exactly is a nearby view is a difficult question, and the number of nearby views we consider varies greatly with the viewpoint. We refer the reader to [2] for more details. This procedure typically results in several thousand selected elements for each architectural site. Examples of the final visual elements obtained by the proposed approach are shown in figure 4.

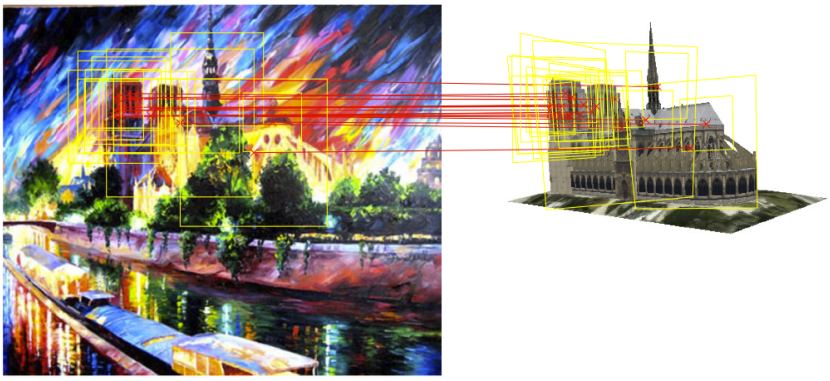


Fig. 5 Illustration of alignment. We use the recovered discriminative visual elements to find correspondences between the input scene depiction (left) and a geo-referenced 3D model (right). Shown is the recovered viewpoint and inlier visual elements found via RANSAC.

3.3 Geo-localization by recovering viewpoint

In this section we describe how, given the set of discriminative visual elements gleaned from the set of 3D models, we identify which 3D site is depicted in the input depiction, and recover the viewpoint of the input depiction with respect to the 3D model. We assume that the depictions are perspective scene renderings and seek to recover the camera center and the camera rotation matrix via camera resectioning [27]. As all 3D models are georeferenced, the recovered camera position and viewpoint provide a geo-localization of the input depiction.

For detection, each discriminative visual element takes as input a 2D patch from the depiction and returns as output a 3D model ID, a 3D location $\mathbf{X}$ on the 3D model, a plane representing the patch extent on the 3D model centered at $\mathbf{X}$, and a detector response score indicating the quality of the appearance match. Following the matching procedure described in section 3.2.4, we form a set of putative discriminative visual element matches using the following procedure. First, we apply to the input depiction all visual element detectors from all 3D models and take the top 200 detections sorted according to the first to second nearest neighbor ratio [35], using the calibrated similarity score (9). This selects the least ambiguous matches. Second, we sort the 200 matches directly by score (9) and select the top 25 matches. This two step selection process choses putative matches that are both non-ambiguous (step 1) and have a high matching score (step 2). From each putative visual element match we obtain 5 putative point correspondences by taking the 2D/3D locations of the patch center and its four corners. The patch corners provide information about the patch scale and the plane location on the 3D model, which has been shown to work well for structure-from-motion with planar constraints [49]. At this point the putative correspondences could still match to several different 3D models. To resolve this, we use RANSAC [19] to find the set of inlier correspondences to a camera model with a constraint that inliers must come from the same architectural site.

Table 1 Statistics of the dataset of historical photographs and non-photographic depictions used for our quantitative evaluation.

	S. Marco Basilica	Trevi Fountain	Notre Dame	**Total**
Hist. photos	30	0	41	**71**
Paintings	41	34	52	**127**
Drawings	19	5	34	**58**
Engravings	9	10	20	**39**
Total	**99**	**49**	**147**	**295**

Among the 125 putative correspondences derived from the 25 putative matches, at each RANSAC iteration we first select a correspondence at random, followed by a random selection of two correspondences from the same 3D model. We use the three points to estimate a camera matrix and then compute the number of inlier correspondences to the camera matrix. We use a restricted camera model where the intrinsics are fixed with the focal length set to the image diagonal length and the principal point set to the center of the image. The result of this RANSAC procedure is both a best matching 3D model and the corresponding camera matrix. The recovered viewpoint geo-localizes the input depiction as well as provides an alignment of the input depiction to the 3D model, as shown in figure 5.

4 Results

In this section we evaluate the potential of our method for geo-localizing a given 2D depiction across different architectural sites and the quality of the recovered viewpoint. A detailed analysis of the 2D-3D instance alignment pipeline, as well as comparisons with other methods are given in [2]. Note that this prior work aligns a historical photograph or a non-photographic depiction to a 3D model of the depicted site assuming the identity of the depicted site is known. Here we are interested in the harder task of identifying the correct architectural site among a set of given 3D models of different architectural sites.

In the following, dataset and performance measures are described in section 4.1, quantitative evaluation is given in section 4.2 and qualitative results are shown in section 4.3. Finally, the main failure modes are discussed in section 4.4.

4.1 Dataset and performance measures

We consider a subset of the dataset introduced in [2] consisting of 3D models and historical photographs/non-photographic depictions of three architectural landmarks. The dataset contains 3D models downloaded from Trimble 3D Warehouse for the following architectural landmarks: Notre Dame of Paris, Trevi Fountain, and San Marco's Basilica. The 3D models consist of basic primitive shapes and have

Table 2 The percentage of input depictions that were assigned to the correct architectural site split across different sites (rows) and depiction styles (columns). Note that there are no historical photographs for Trevi Fountain in the database of [2].

	Paintings	Historical photograph	Engravings	Drawings	**Average**
S. Marco Basilica	83%	87%	89%	94%	**87%**
Trevi Fountain	82%	-	90%	80%	**84%**
Notre Dame	90%	88%	85%	79%	**86%**
Average	**86 %**	**87%**	**87%**	**84%**	**86%**

a composite texture from a set of images. The 2D depictions for the three sites were collected by [2] from the Internet and include 71 historical photographs and 224 non-photographic depictions, with 39 engravings, 58 drawings, and 127 paintings. The drawings category includes color renderings and the paintings category includes different rendering styles, such as watercolors, oil paintings, and pastels. Table 1 shows the number of images belonging to each category across the different sites.

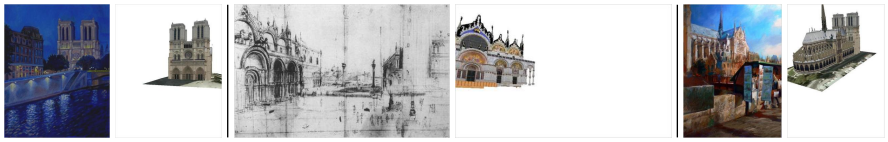
We measure performance for the following two tasks. First, we evaluate the geo-localization accuracy, which measures the percentage of input depictions that are matched to the correct architectural site. Second, for the depictions assigned to the correct architectural site we evaluate the quality of the resulting alignment, which is measured by a user study via Amazon Mechanical Turk.

4.2 Quantitative evaluation

We summarized the three Trimble 3D Warehouse models with 15,000 discriminative visual elements each. For each input depiction, we applied all of the 45,000 detectors corresponding to those elements, selected the 25 most confident ones, and performed camera resectioning using RANSAC as described in section 3.3, with the constraint that only elements from the same site could be counted as inliers. Thus, our output is both a specific 3D model and a viewpoint.

We first report results on the task of identifying the 3D model of the architectural site. Table 2 shows the results separately for the three different sites and across different depiction styles. Despite the difficulty of the task due to the large variety of viewpoints and styles, our method identified correctly the architectural site for 86% of the depictions, which is much larger than the 33% chance performance.

We then evaluated the quality of the alignments for depictions that were assigned to the correct site. To quantitatively evaluate the goodness of our alignments, we have conducted a user study via Amazon Mechanical Turk. As in [2], the workers were asked to judge the viewpoint similarity of the resulting alignments to their corresponding input depictions by categorizing the viewpoint similarity as either a (a) Good match, (b) Coarse match, or (c) No match, illustrated in figure 6. We asked



(a) Good match (b) Coarse match (c) No match

Fig. 6 Alignment evaluation criteria. We asked workers on Amazon Mechanical Turk to judge the viewpoint similarity of the resulting alignment to the input depiction. The workers were asked to categorize the viewpoint similarity into one of three categories: (a) Good match – the two images show a roughly similar view of the building; (b) Coarse match – the view may not be similar, but the building is roughly at the same location in both images, not upside down, and corresponding building parts can be clearly identified; (c) No match – the views are completely different, e.g. upside down, little or no visual overlap.

Fig. 7 Alignment of historical photographs of San Marco's Square (top) and Notre Dame of Paris (bottom) to their respective 3D models.

five different workers to rate the viewpoint similarity for each depiction and we report the majority opinion. Table 3 shows the performance of our algorithm for the different depiction styles. The performance varies across depiction style from 77% of coarse/good matches for paintings to more than 90% for historical photographs or engravings. Overall, 83% of the input depictions are at least coarsely matched.

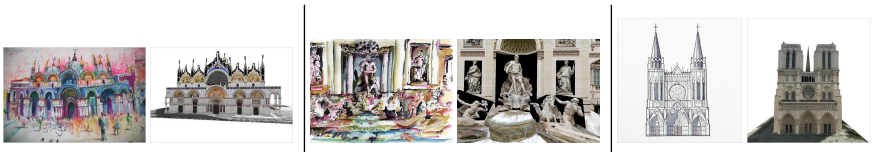
Table 3 Viewpoint similarity user study of our algorithm across different depiction styles.

	Good match	Coarse match	No match
Historical photographs	74%	16%	10%
Paintings	57%	20%	23%
Drawings	59%	20%	20%
Engravings	65%	29%	6%
Average	**63%**	**20%**	**17%**

Fig. 8 Example alignments of non-photographic depictions to 3D models. Notice that we are able to align depictions rendered in different styles and having a variety of viewpoints with respect to the 3D models.

4.3 Qualitative evaluation

Figures 7 and 8 show example alignments of historical photographs and non-photographic depictions, respectively. Notice that the depictions are reasonably well aligned with the 3D models, with regions on the 3D model rendered onto the corresponding location for a given depiction. We are able to cope with a variety of viewpoints with respect to the 3D model as well as different depiction styles. Our approach succeeds in recovering the approximate viewpoint in spite of these challenging appearance changes and the varying quality of the 3D models.



(a) Scene distortion. (b) Drawing and 3D errors. (c) Major structural differences.

Fig. 9 Challenging examples successfully aligned by our method where the assumption of a perspective scene rendering is violated. Note that the drawing in (c) is a completely different cathedral.

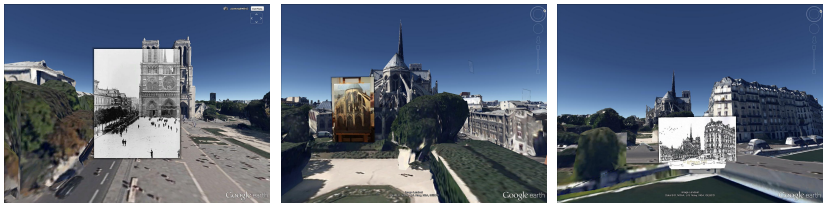
(a) San Marco's Basilica. (b) Trevi Fountain.

Fig. 10 Recovered viewpoints of some of the geo-localized depictions visualized in Google Earth.

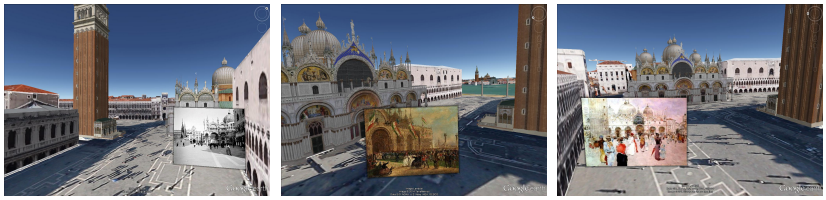
Note that figures 7 and 8, in addition to the Trimble 3D Warehouse models, also include alignment results using a 3D model of San Marco's Square that was reconstructed from a set of photographs using dense multi-view stereo [21]. While the latter 3D model has more accurate geometry than the Trimble 3D Warehouse models, it is also much noisier along the model boundaries. This model was excluded from the quantitative evaluation in section 4.2 as it overlaps with the San Marco Basilica Trimble 3D Warehouse model, but we include it here to demonstrate alignment for different types of 3D models.

In figure 9 we show alignments for a set of challenging examples where the assumption of a perspective rendering is significantly violated, but the proposed approach was still able to recover a reasonable alignment. Notice the severe non-perspective scene distortions, drawing errors, and major architectural differences (e.g. a part of the landmark may take a completely different shape).

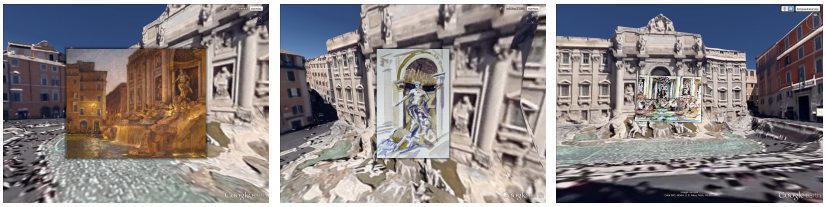
Figures 1 and 10 show the recovered viewpoints of several different depictions for the three sites rendered in Google Earth. Figure 11 shows individual depictions rendered in Google Earth, which showcases a re-photography application by allowing a user to browse the depictions in the context of their modern environments. Please see additional qualitative results on the project webpage [1].



(a) Notre Dame



(b) San Marco's Basilica.



(c) Trevi Fountain.

Fig. 11 Examples of geo-localized depictions visualized in Google Earth. Note that the proposed method allows us to visualize the specific place across time and through the eyes of different artists.

4.4 Failure modes

We have identified three main failure modes of our algorithm, examples of which are shown in figure 12. The first is due to large-scale symmetries, for example when the front and side facade of a building are very similar. This problem is difficult to resolve with only local reasoning. For example, the proposed cross-validation step removes repetitive structures visible in the same view but not at different locations of the site. The second failure mode is due to locally confusing image structures, for example, the vertical support structures on the cathedral in figure 12 (middle) are locally similar (by their HOG descriptor) to the vertical pencil strokes on the drawing. The learnt mid-level visual elements have a larger support than typical local invariant features (such as SIFT) and hence are typically more distinctive. Nevertheless, such mismatches can occur and in some cases are geometrically consistent with a certain view of the 3D model. The third failure mode is when the depicted viewpoint is not covered in the set of sampled views. This can happen for unusual viewpoints including extreme angles, large close-ups, or cropped views. Such unusual views are in some cases assigned to a wrong 3D site.

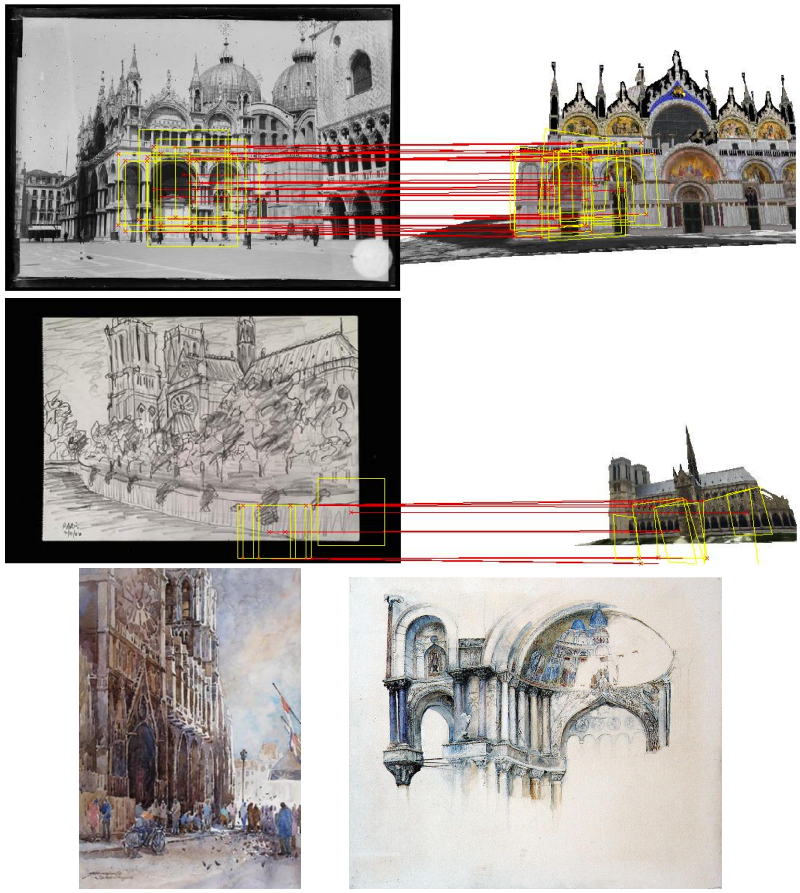


Fig. 12 Example failure cases. Top: large scale symmetry. Here arches are incorrectly matched on a building with similar front and side facades. Middle: locally confusing image structures. Here the vertical support structures on the cathedral (right) are locally similar by their HOG descriptor to the vertical pencil strokes on the drawing (left). Bottom: Two examples of paintings with unusual viewpoints.

5 Conclusion

We have demonstrated that automatic geo-localization is possible for a range of non-photographic depictions and historical photographs, which represent extremely challenging cases for current local feature matching methods. To achieve this we have developed an approach to compactly represent 3D models of architectural sites by a set of visually distinct mid-level scene elements extracted from rendered views, and have shown that they can be reliably matched in a variety of photographic and non-photographic depictions. We have also shown an application of the proposed approach to computational re-photography to automatically geo-tag and find

an approximate viewpoint of historical photographs and paintings, which allows for geo-browsing within Google Earth. This work is just a step towards computational reasoning about the content of non-photographic depictions. The developed approach for extracting visual elements opens-up the possibility of efficient indexing for visual search of paintings and historical photographs (e.g. via hashing of the HOG features as in [14]), or automatic fitting of complex non-perspective models used in historical imagery [40]. It would be also interesting to investigate learning our 3D mid-level visual elements with convolutional neural network descriptors [16, 24, 32, 38, 44, 52], which have recently shown promising results in object detection in non-photographic depictions [10].

Acknowledgements

We are grateful to Guillaume Seguin, Alyosha Efros, Guillaume Obozinski and Jean Ponce for their useful feedback, and to Yasutaka Furukawa for providing access to the San Marco 3D model. This work was partly supported by the EIT ICT Labs, ANR project SEMAPOLIS (ANR-13-CORD-0003) and the ERC starting grant LEAP. The work was partly carried out at IMAGINE, a joint research project between Ecole des Ponts ParisTech (ENPC) and the Scientific and Technical Centre for Building (CSTB). Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory. The U.S. Government is auhorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL or the U.S. Government.

References

1. Http://www.di.ens.fr/willow/research/painting_to_3d/
2. Aubry, M., Russell, B., Sivic, J.: Painting-to-3D model alignment via discriminative visual elements. ACM Transactions on Graphics **33**(2) (2014)
3. Baatz, G., Saurer, O., Köser, K., Pollefeys, M.: Large scale visual geo-localization of images in mountainous terrain. In: Proceedings of European Conference on Computer Vision (2012)
4. Baboud, L., Cadik, M., Eisemann, E., Seidel, H.P.: Automatic photo-to-terrain alignment for the annotation of mountain pictures. In: Proceedings of the conference on Computer Vision and Pattern Recognition (2011)
5. Bach, F., Harchaoui, Z.: Diffrac : a discriminative and flexible framework for clustering. In: Advances in Neural Information Processing Systems (2008)
6. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
7. Bosché, F.: Automated recognition of 3D CAD model objects in laser scans and calculation of as-built dimensions for dimensional compliance control in construction. Advanced engineering informatics **24**(1), 107–118 (2010)
8. Chen, D., Baatz, G., et al.: City-scale landmark identification on mobile devices. In: Proceedings of the conference on Computer Vision and Pattern Recognition (2011)
9. Chum, O., Matas, J.: Geometric hashing with local affine frames. In: Proceedings of the conference on Computer Vision and Pattern Recognition (2006)

10. Crowley, E.J., Zisserman, A.: In search of art. In: Workshop on Computer Vision for Art Analysis, ECCV (2014)
11. Crowley, E.J., Zisserman, A.: The state of the art: Object retrieval in paintings using discriminative regions. In: British Machine Vision Conference (2014)
12. Cummins, M., Newman, P.: Highly scalable appearance-only SLAM - FAB-MAP 2.0. In: Proceedings of Robotics: Science and Systems. Seattle, USA (2009)
13. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: Proceedings of the conference on Computer Vision and Pattern Recognition (2005)
14. Dean, T., Ruzon, M., Segal, M., Shlens, J., Vijayanarasimhan, S., Yagnik, J.: Fast, accurate detection of 100,000 object classes on a single machine. In: Proceedings of the conference on Computer Vision and Pattern Recognition (2013)
15. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.A.: What makes paris look like paris? ACM Transactions on Graphics (Proc. SIGGRAPH) **31**(4) (2012)
16. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. arXiv:1310.1531 (2013)
17. Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: Liblinear: A library for large linear classification. Journal of Machine Learning Research **9**(1), 1871–1874 (2008)
18. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions Pattern Analysis and Machine Intelligence **32**(9) (2010)
19. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
20. Frome, A., Singer, Y., Sha, F., Malik, J.: Learning globally-consistent local distance functions for shape-based image retrieval and classification. In: Proceedings of International Conference on Computer Vision (2007)
21. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Towards internet-scale multi-view stereo. In: Proceedings of the conference on Computer Vision and Pattern Recognition (2010)
22. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multi-view stereopsis. IEEE Transactions Pattern Analysis and Machine Intelligence **32**(8) (2010)
23. Gharbi, M., Malisiewicz, T., Paris, S., Durand, F.: A Gaussian approximation of feature space for fast image similarity. Tech. rep., MIT (2012)
24. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the conference on Computer Vision and Pattern Recognition (2014)
25. Gronat, P., Obozinski, G., Sivic, J., Pajdla, T.: Learning and calibrating per-location classifiers for visual place recognition. In: Proceedings of the conference on Computer Vision and Pattern Recognition (2013)
26. Hariharan, B., Malik, J., Ramanan, D.: Discriminative decorrelation for clustering and classification. In: Proceedings of European Conference on Computer Vision (2012)
27. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, second edn. Cambridge University Press, ISBN: 0521540518 (2004)
28. Hauagge, D., Snavely, N.: Image matching using local symmetry features. In: Proceedings of the conference on Computer Vision and Pattern Recognition (2012)
29. Huttenlocher, D.P., Ullman, S.: Object recognition using alignment. In: International Conference on Computer Vision (1987)
30. Irschara, A., Zach, C., Frahm, J.M., Bischof, H.: From structure-from-motion point clouds to fast location recognition. In: Proceedings of the conference on Computer Vision and Pattern Recognition (2009)
31. Knopp, J., Sivic, J., Pajdla, T.: Avoiding confusing features in place recognition. In: Proceedings of European Conference on Computer Vision (2010)
32. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
33. Li, Y., Snavely, N., Huttenlocher, D., Fua, P.: Worldwide pose estimation using 3D point clouds. In: Proceedings of European Conference on Computer Vision (2012)

34. Lowe, D.: The viewpoint consistency constraint. International Journal of Computer Vision **1**(1), 57–72 (1987)
35. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
36. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-svms for object detection and beyond. In: Proceedings of International Conference on Computer Vision (2011)
37. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: Proceedings of the conference on Computer Vision and Pattern Recognition (2006)
38. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)
39. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proceedings of the conference on Computer Vision and Pattern Recognition (2007)
40. Rapp, J.: A geometrical analysis of multiple viewpoint perspective in the work of Giovanni Battista Piranesi: an application of geometric restitution of perspective. The Journal of Architecture **13**(6) (2008)
41. Russell, B.C., Sivic, J., Ponce, J., Dessales, H.: Automatic alignment of paintings and photographs depicting a 3D scene. In: IEEE Workshop on 3D Representation for Recognition (3dRR-11), associated with ICCV (2011)
42. Sattler, T., Leibe, B., Kobbelt, L.: Fast image-based localization using direct 2d-to-3d matching. In: Proceedings of International Conference on Computer Vision (2011)
43. Schindler, G., Brown, M., Szeliski, R.: City-scale location recognition. In: Proceedings of the conference on Computer Vision and Pattern Recognition (2007)
44. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv:1312.6229 (2013)
45. Shalev-Shwartz, S., Singer, Y., Srebro, N., Cotter, A.: Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. Mathematical Programming, Series B **127**(1), 3–30 (2011)
46. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: Proceedings of the conference on Computer Vision and Pattern Recognition (2007)
47. Shrivastava, A., Malisiewicz, T., Gupta, A., Efros, A.A.: Data-driven visual similarity for cross-domain image matching. In: ACM Transactions on Graphics (Proc. SIGGRAPH Asia) (2011)
48. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proceedings of International Conference on Computer Vision (2003)
49. Szeliski, R., Torr, P.: Geometrically constrained structure from motion: Points on planes. In: European Workshop on 3D Structure from Multiple Images of Large-Scale Environments (SMILE) (1998)
50. Torii, A., Sivic, J., Pajdla, T., Okutomi, M.: Visual place recognition with repetitive structures. In: Proceedings of the conference on Computer Vision and Pattern Recognition (2013)
51. Zamir, A., Shah, M.: Accurate image localization based on google maps street view. In: Proceedings of European Conference on Computer Vision (2010)
52. Zeiler, M., Fergus, R.: Visualizing and understanding convolutional networks. arXiv:1311.2901 (2013)